Grid File Replication using Storage Resource Management

Presented By Alex Sim

Contributors:

JLAB: Bryan Hess, Andy Kowalski

Fermi: Don Petravick, Timur Perelmutov, Rich Wellner

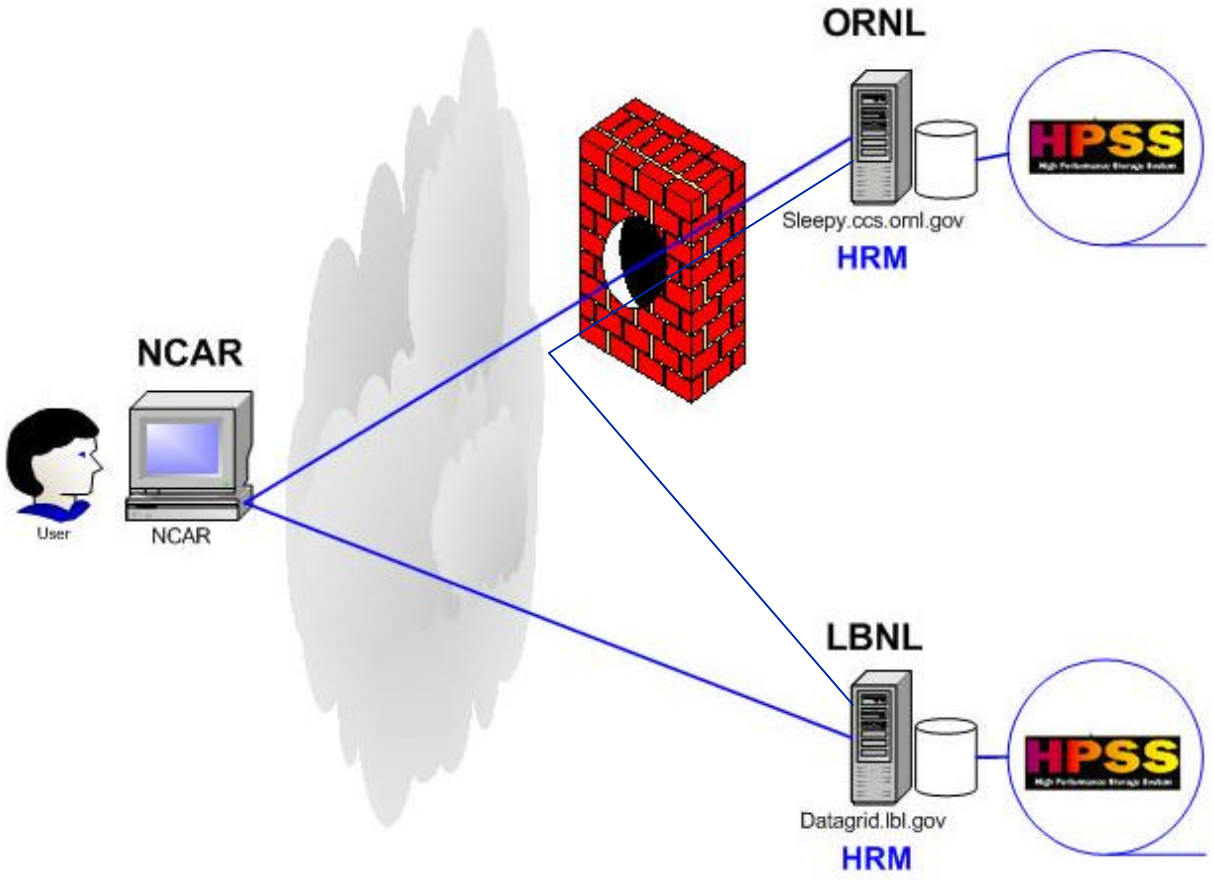
LBNL: Junmin Gu, Vijaya Natarayan, Ekow Otoo, Alex Romosan, Alex Sim, Arie Shoshani

WP2-EDG: Wolfgang Hoschek, Peter Kunszt, Heinz Stockinger, Kurt Stockinger

WP5-EDG: Jean-Philippe Baud

http://sdm.lbl.gov/srm

How does file replication use SRMs
(high level view)



Main advantages of using SRMs for file replication

- **Can work in front of MSS**
  - To provide pre-staging
  - To provide queued archiving
- **Monitor transfer in/out of MSS**
  - Recover in case of transient failures
  - Reorders pre-staging requests to minimize tape mounts
- **Monitors GridFTP transfers**
  - Re-issues requests in case of failure
- **Can control number of concurrent GridFTP transfers to optimize network use (future)**
- **SRMs role in the data replication**
  - Storage resource coordination & scheduling
  - SRMs do not perform file transfer
  - SRMs do invoke file transfer service if needed (GridFTP)

Brief history of SRM since GGF4

- **Agreed on single API for multiple storage systems**
  - **Jlab has an SRM implementation based on SRM v1.1 spec on top of JASMine**
  - **Fermi Lab has an SRM implementation based on SRM v1.1 spec on top of Enstore**
  - **WP5-EDG is proceeding with SRM implementation on top of Castor**
  - **LBNL Deployed HRM-HPSS (which accesses files in/out of HPSS) at BNL, ORNL, and NERSC (PDSF)**
  - **HRM-NCAR that accesses MSS at NCAR is in progress**

SRM version 2.0

- Joint design and specification of SRM v2.0
  - LBNL organized meeting to coordinate design, and summarized design conclusions
    - SRM v2.0 spec draft version exists
    - SRM v2.0 finalization to be done in Dec.
- Design uses OGSA service concept
  - Define interface & behavior
  - Select protocol binding (WSDL/SOAP)
  - Permit multiple implementations
    - Disk Resource Managers (DRMs)
    - On top of multiple MSSs (HRMs)

Brief Summary
SRM main methods

- **srmGet, srmPut, srmCopy**
  - Multiple files
  - srmGet from remote location to disk/tape
  - srmPut from client to SRM disk/tape
  - srmCopy from remote location to SRM disk/tape
- **srmRelease**
  - Pinning automatic
  - If not provided, apply pinning lifetime
- **srmStatus**
  - Per file, per request
  - Time estimate
- **srmAbort**

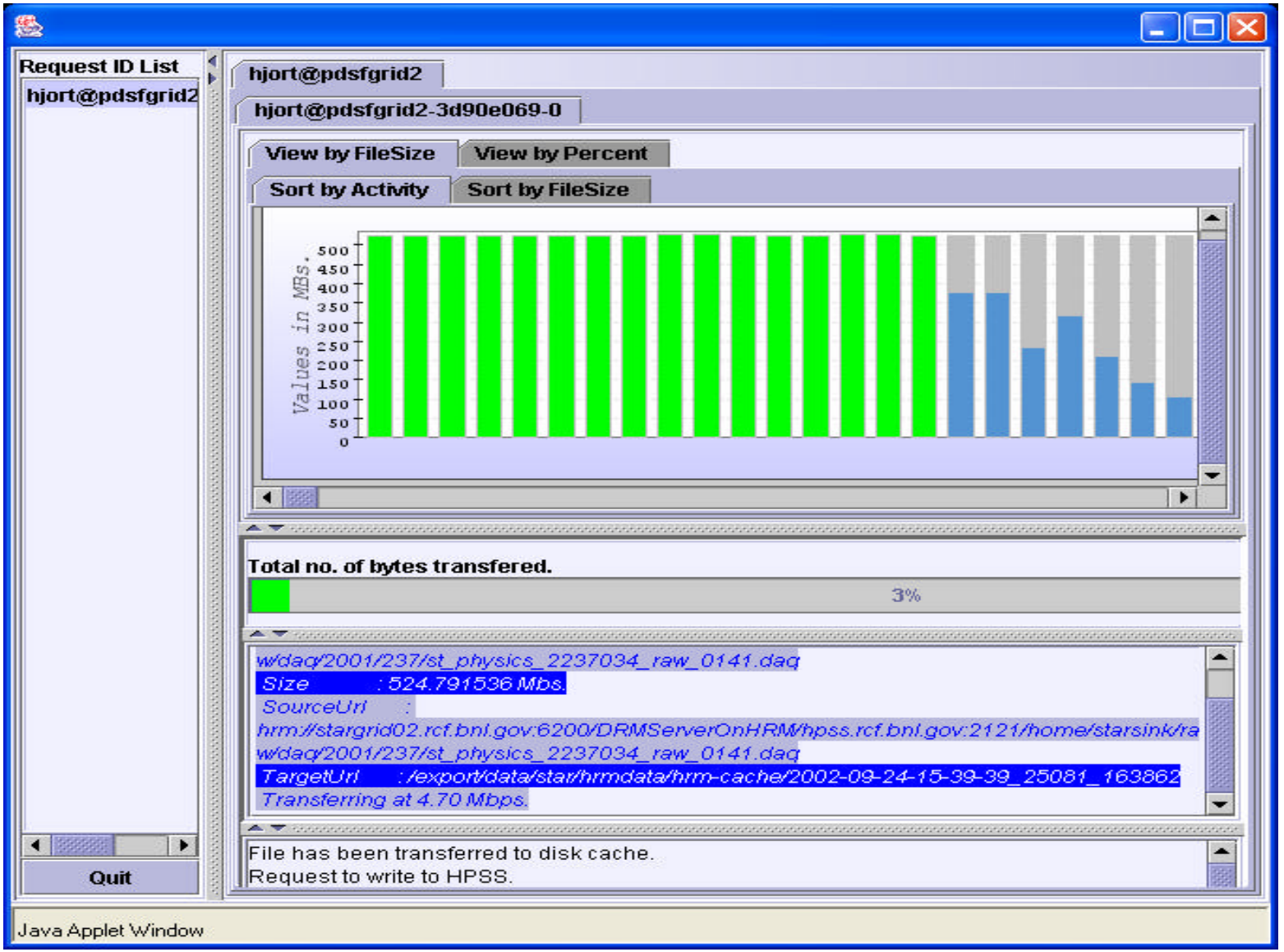
Main Design Points

- Interfaces to all types of SRMs to be uniform
- Any Clients, Middleware modules, other SRMs
  - Will communicate with SRMs
- Support a “multiple files” request
  - set of files, not ordered, no bundles
  - Implies: queuing, status, time estimates, abort
- SRMs support asynchronous requests
  - Non-blocking, unlike FTP and other services
  - Support long delays, multi-file requests
- Support call-backs
  - Plan to use “event notification service”
- Automatic Garbage Collection
  - In file replication, all files are “volatile”
  - As soon as they are moved to target,
    SRMs perform “garbage collection” automatically.

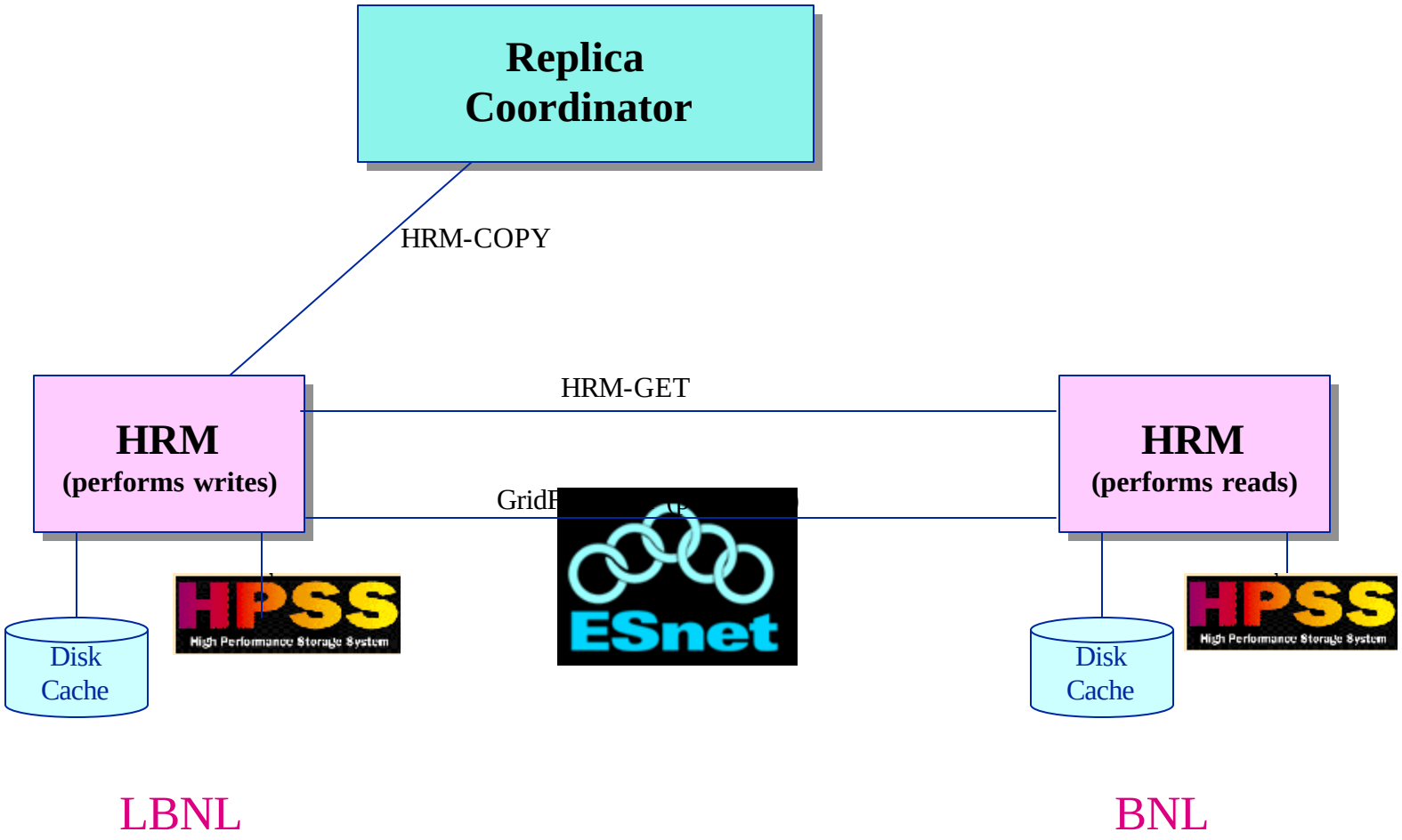
Current efforts on SRMs at LBNL

- Deployed HRM at BNL, LBNL (NERSC/PDSF), ORNL for HPSS access
- Developing HRM at NCAR for MSS access
- Deployed GridFTP-HRM-HPSS connection daemon
- Supports multiple transport protocols
  - gsiftp, ftp, http, bbftp and hrm
- Deployed web-based File Monitoring Tool for HRM
  - especially useful for large file replica requests
- Deployed HRM client command programs
  - User convenience
- Currently developing web-services gateway
- Currently developing GSI-enabled requests

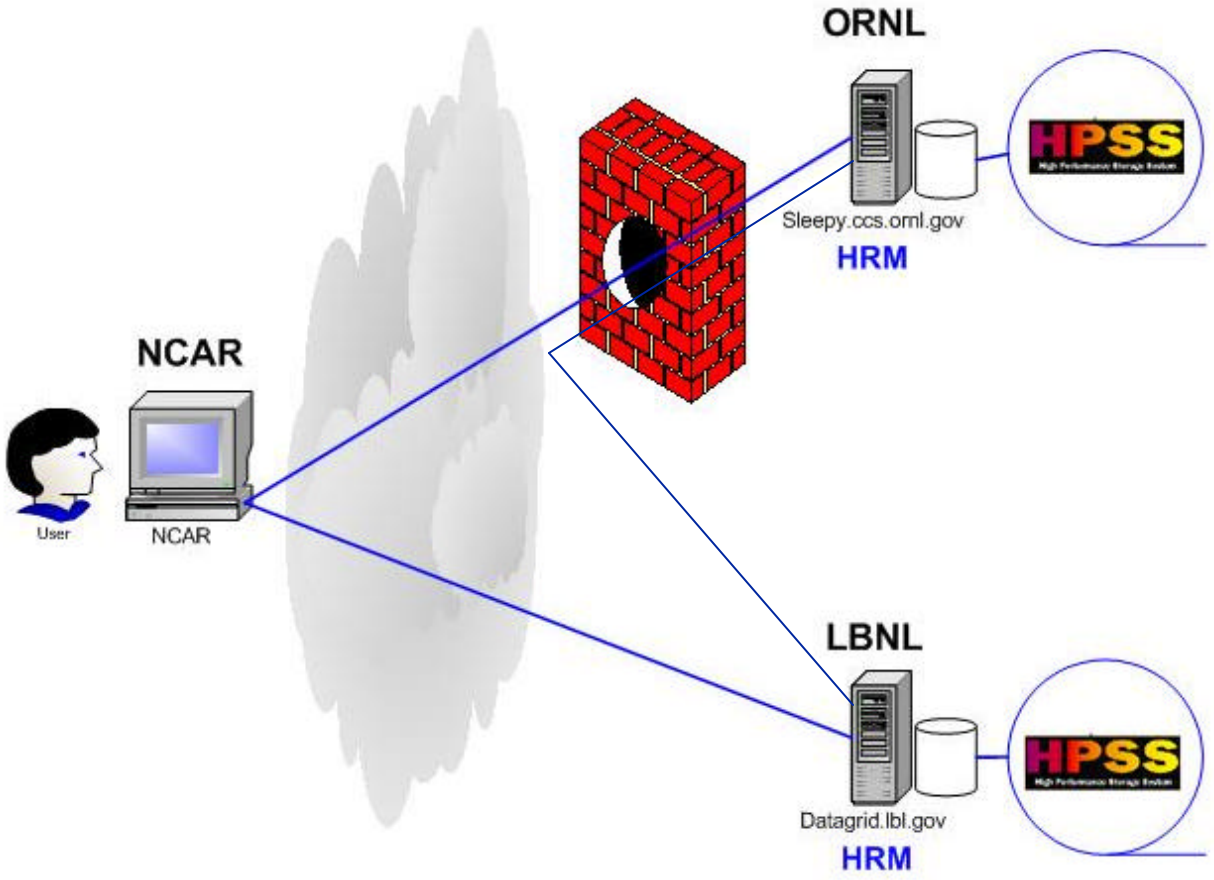
Web-Based File Monitoring Tool



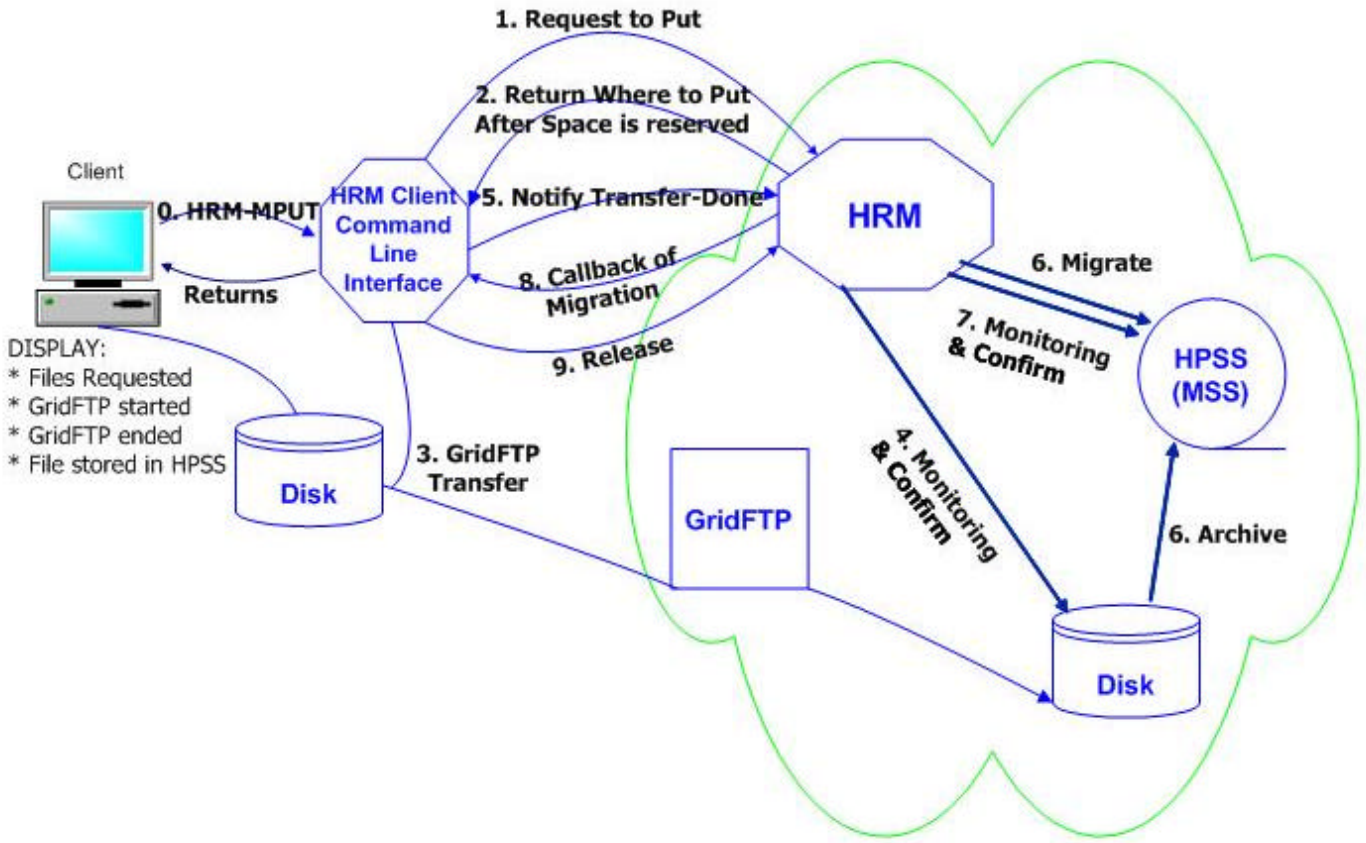
HRMs in PPDG
(high level view)



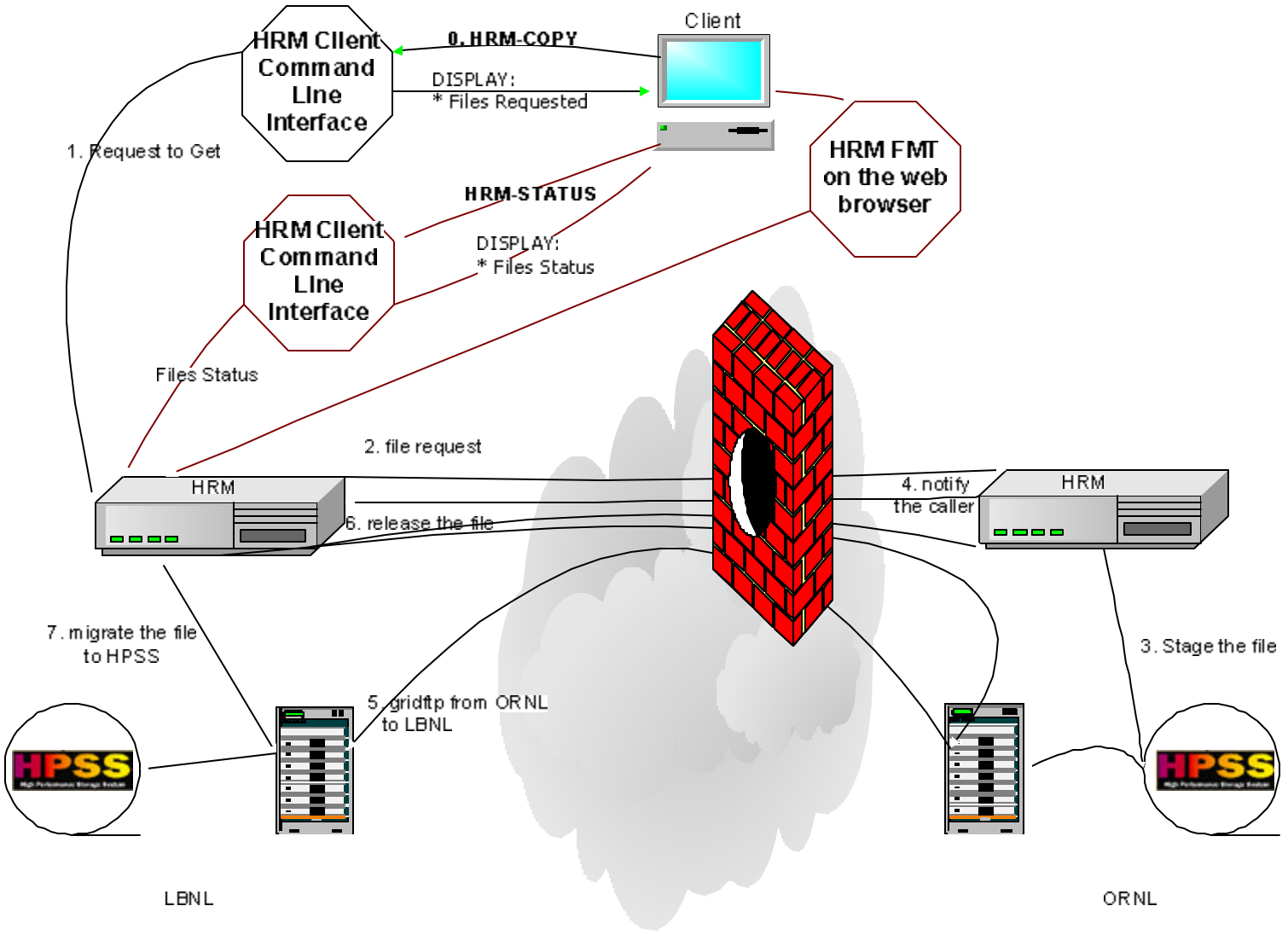
HRMs in ESG
(high level view)



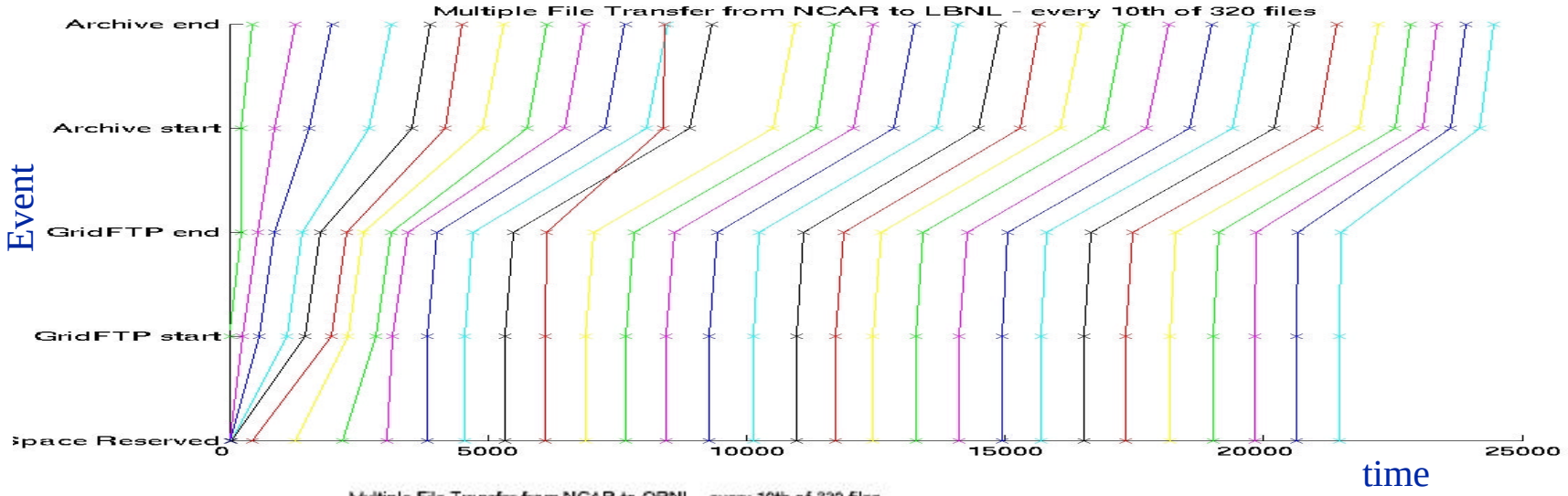
File replication from NCAR to ORNL/LBNL HPSS controlled at NCAR

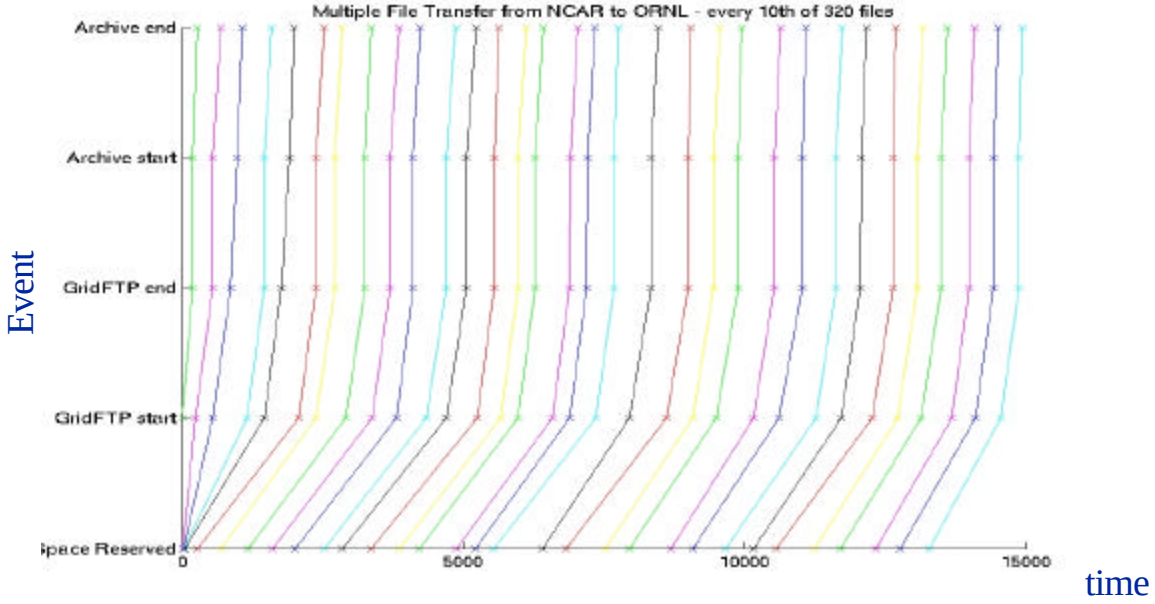


File Replication from ORNL HPSS to NERSC HPSS controlled at NCAR : "Non-Blocking Calls"

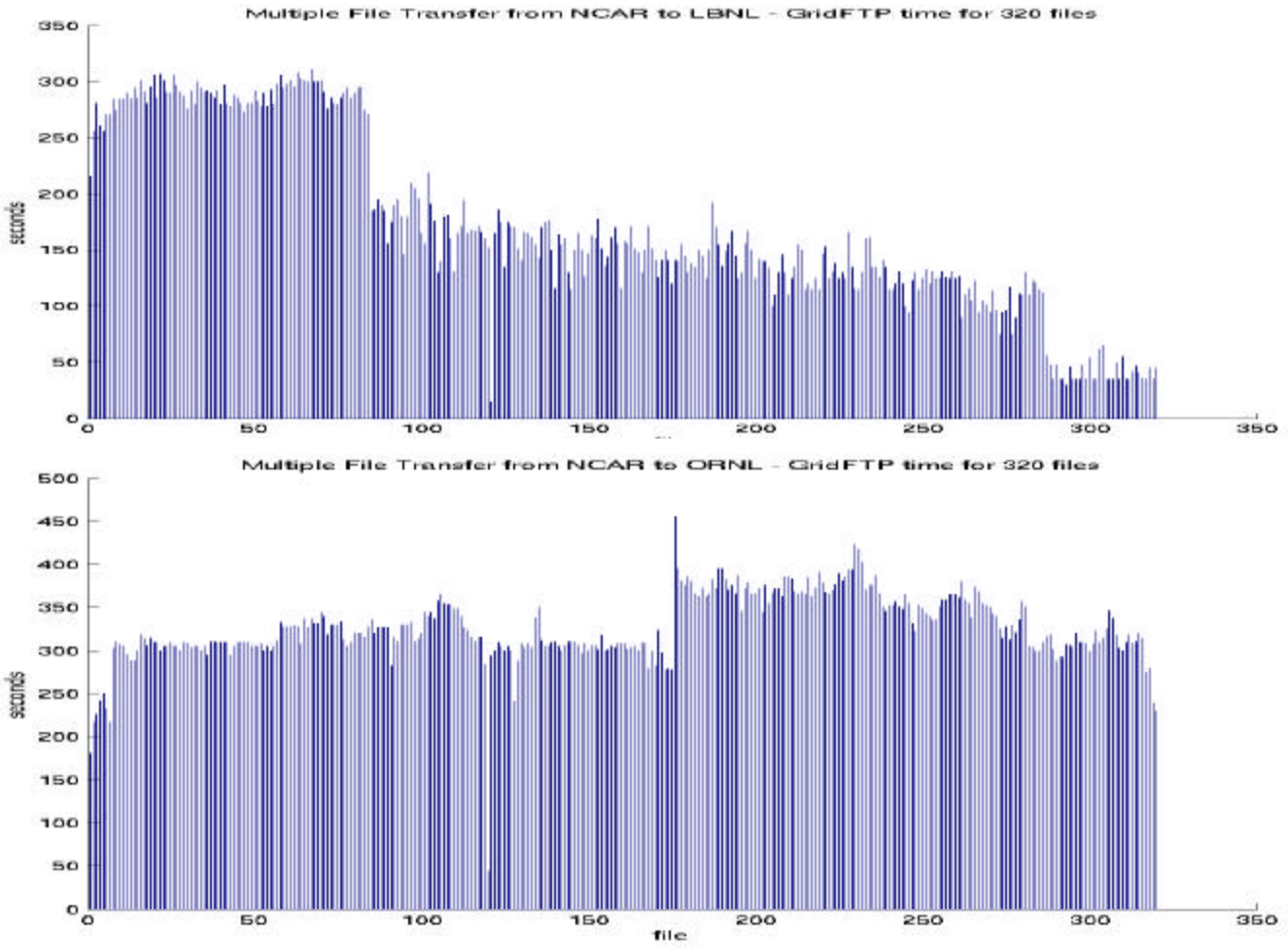


Recent Measurements of large multi-file replication

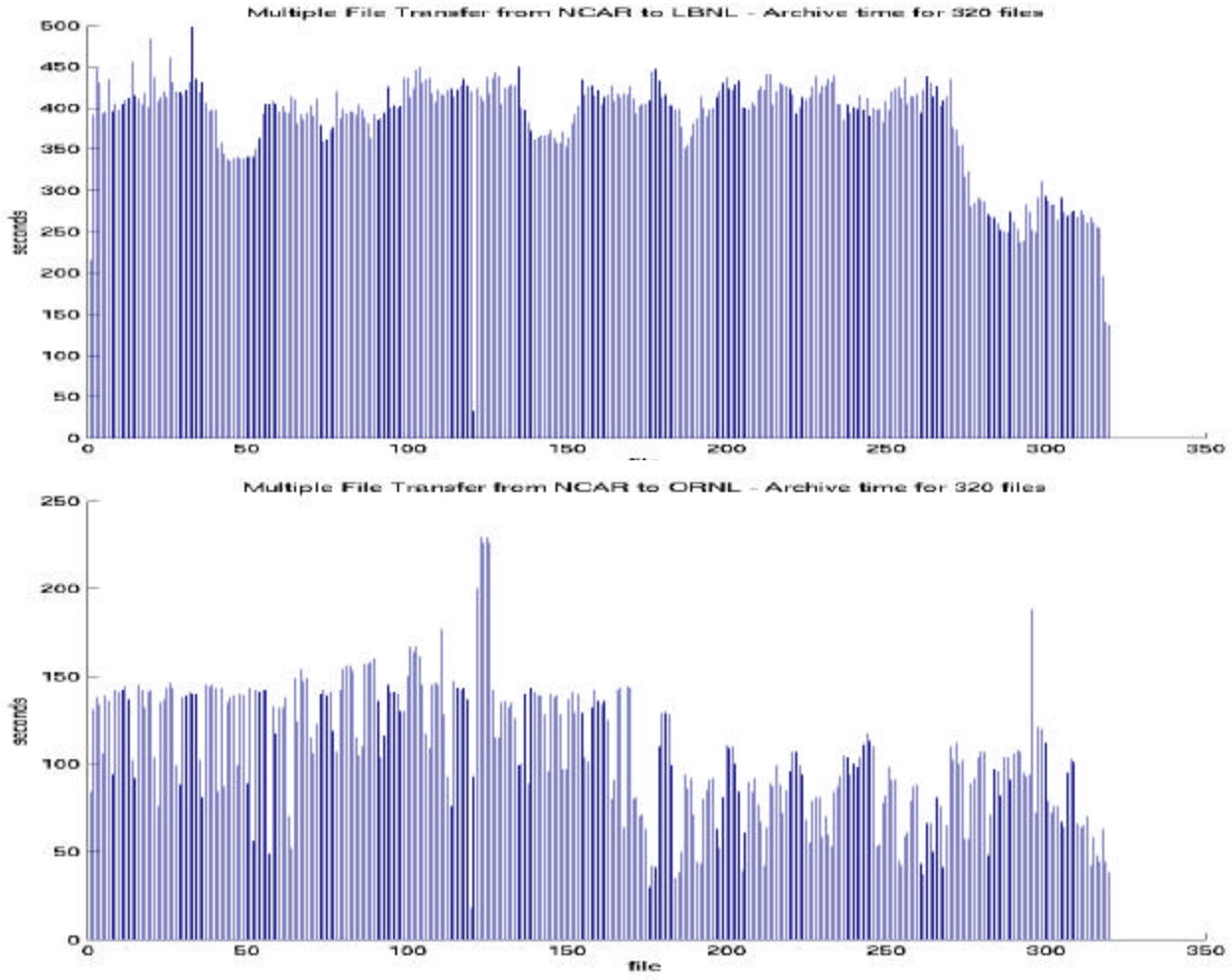




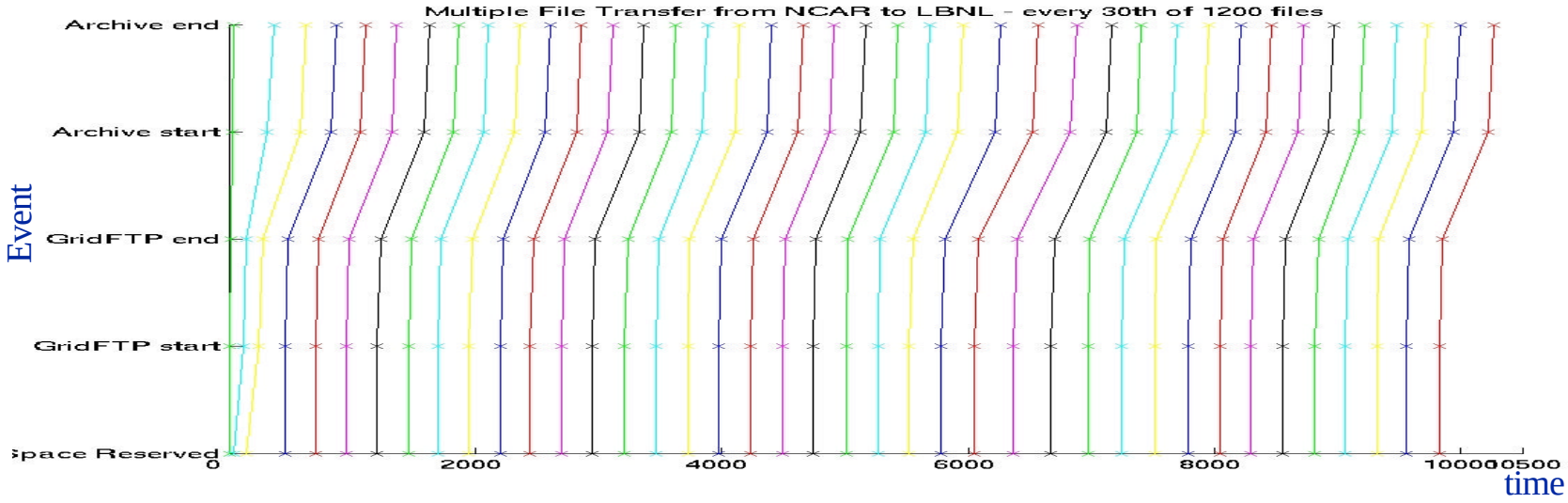
Recent Measurements of large multi-file replication (GridFTP transfer time)

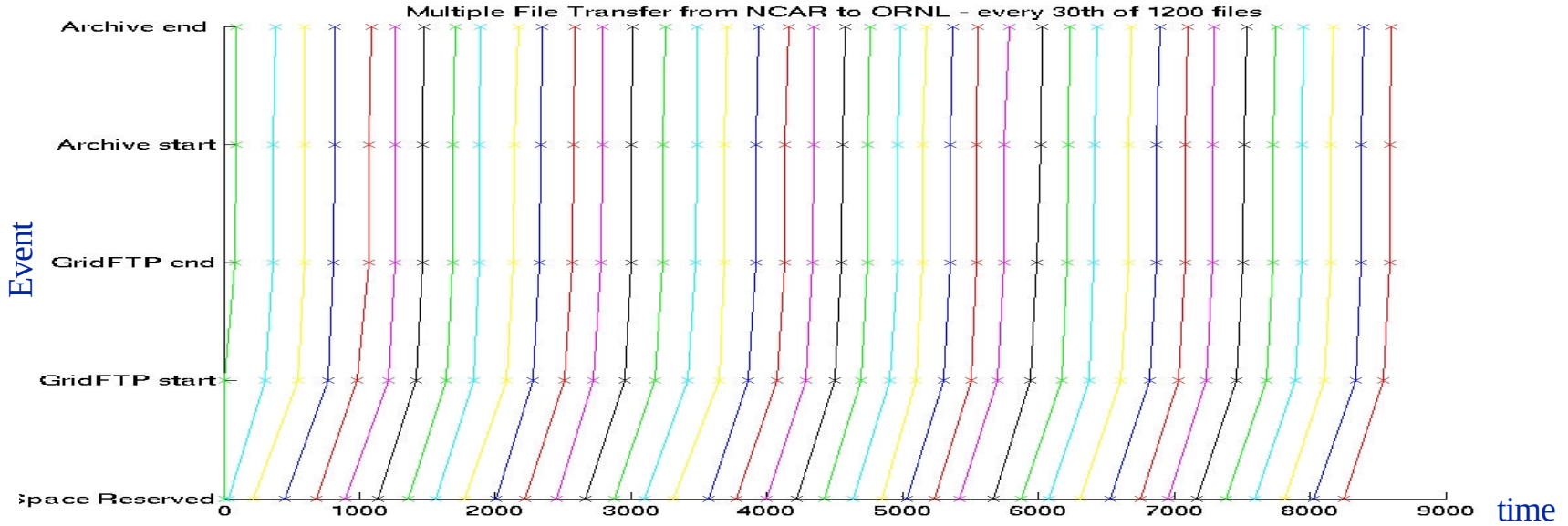


Recent Measurements of large multi-file replication (Archiving time)

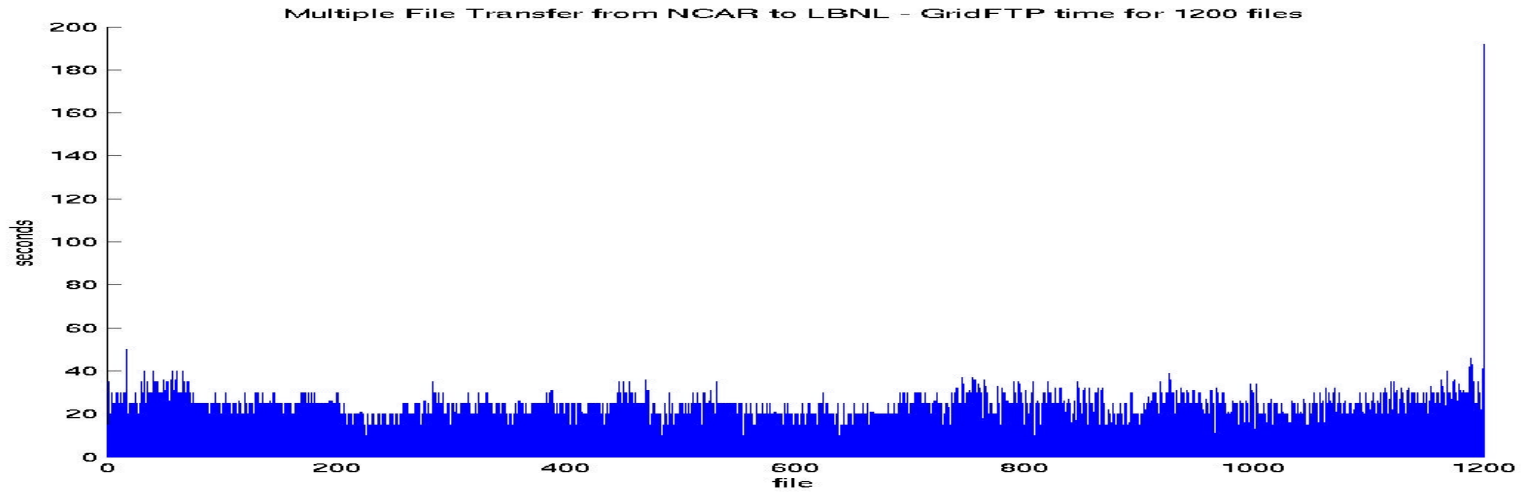


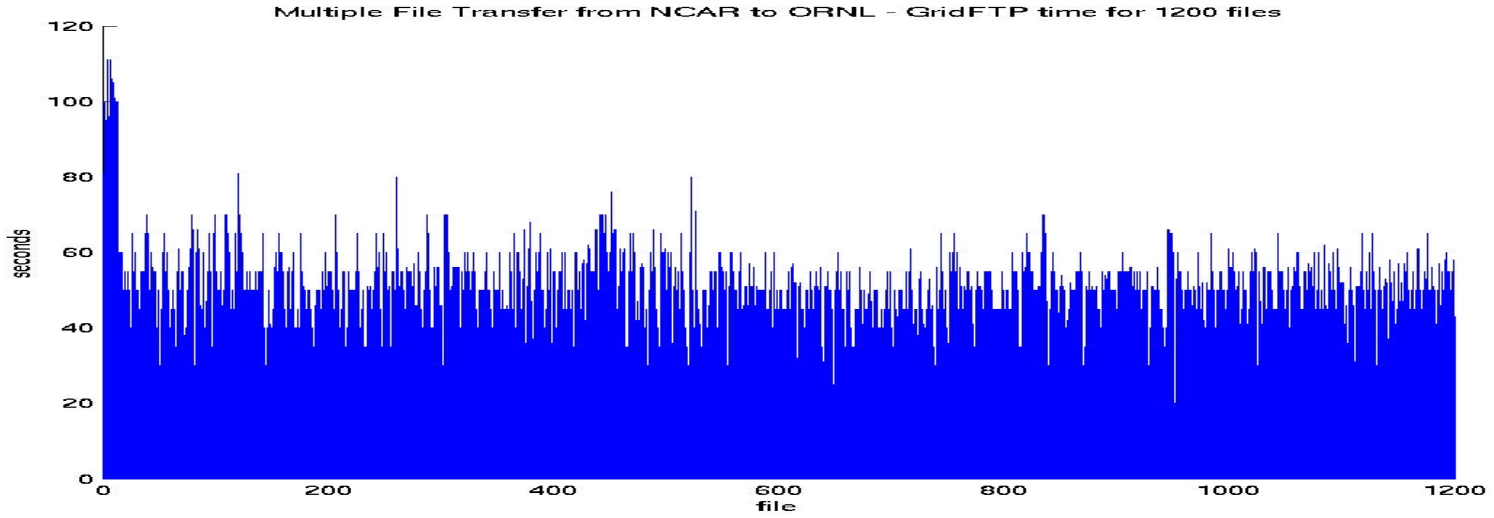
Recent Measurements of large multi-file replication



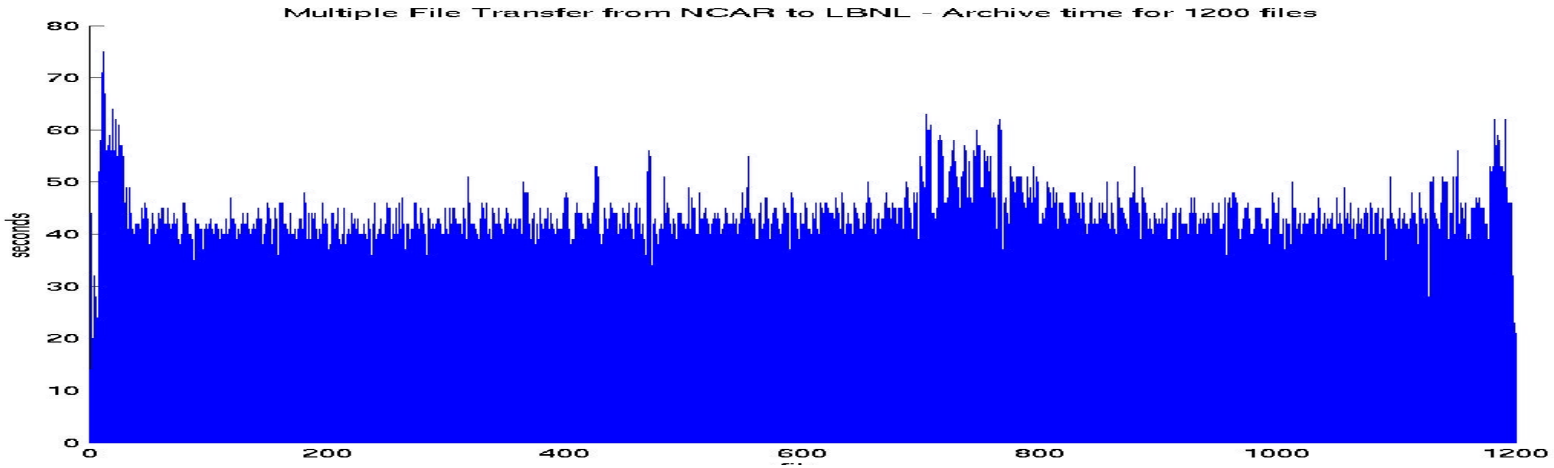


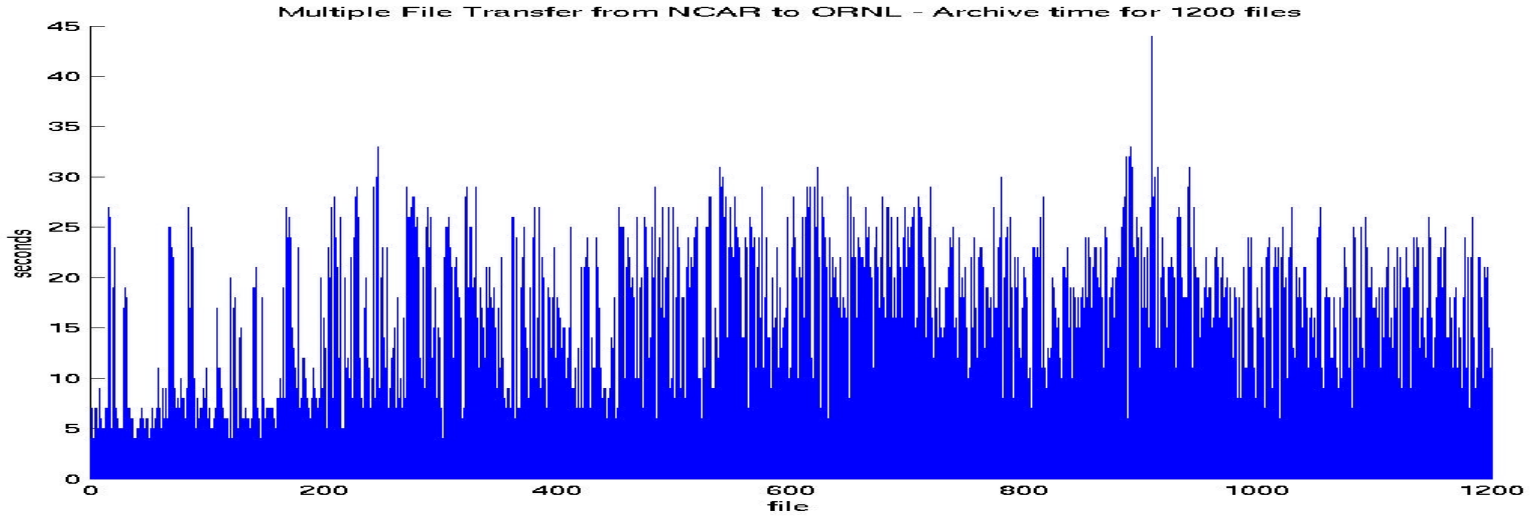
Recent Measurements of large multi-file replication (GridFTP transfer time)





Recent Measurements of large multi-file replication (Archiving time)





File Replication from ORNL HPSS to NERSC HPSS controlled at NCAR